GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Whole-genome duplications contributed to the expansion of cytochrome b_5 genes in *Paramecium tetraurelia*

K.J. Wu*, D. Wang*, J. Ding, S.H. Yang and X.H. Zhang

State Key Laboratory of Pharmaceutical Biotechnology,
Department of Biology, School of Life Sciences, Nanjing University,
Nanjing, China

*These authors contributed equally to this study.
Corresponding authors: X.H. Zhang / S.H. Yang
E-mail: xiaohuizhang@nju.edu.cn / sihaiyang@nju.edu.cn

Genet. Mol. Res. 12 (2): 1882-1896 (2013)
Received February 1, 2012
Accepted July 20, 2012
Published January 9, 2013
DOI http://dx.doi.org/10.4238/2013.January.9.1

ABSTRACT. Cytochrome b_5 (cyt b_5) genes encode ubiquitous electron transport hemoproteins found in animals, plants, fungi, and purple bacteria. However, little is known about their evolutionary history in genomes so far. Here, we conducted an extensive genome-wide survey of cyt b_5 genes in 20 representative model species and identified 310 of these genes. Both the absolute number and relative proportion of cyt b_5 genes in *Paramecium tetraurelia* were significantly higher than those in other genomes. Our data also showed that whole-genome duplications (WGDs), especially the recent WGD, contributed to the species-specific expansion of cyt b_5 genes in the *Paramecium* genome. Furthermore, 24 cyt b_5 genes were identified as the minimal number of ancestral cyt b_5 in the ancestral *Paramecium* genome, which is also the largest number of these genes encountered in an organism. These results suggest that an excess of cyt b_5 genes were selectively retained in this species even before the three WGDs took place. Although more cyt b_5 genes were retained in *P. tetraurelia* than in other genomes, more cyt b_5 losses were

also observed in the *P. tetraurelia* genome, suggesting that the balance of gene retention and loss maintained an optimum dosage of cyt b_5 genes.

Key words: Cytochrome b_5; *Paramecium tetraurelia*; Expansion; Whole-genome duplication; Gene loss

INTRODUCTION

The cytochrome b_5 (cyt b_5) genes constitute a conserved gene family of heme-binding proteins ubiquitously distributed in different life forms, including unicellular eukaryotes (protists and fungi) and multicellular eukaryotes (plants and animals) (Vergeres and Waskell, 1995; Soucy and Luu-The, 2002; Ogishima et al., 2003). Cyt b_5 is found as a membrane component or as a soluble form in cells. Its heme-binding region consists of 6 α-helices and 5 β-strands (β1-α1-β4-β3-α2-α3-β5-α4-α5-β2-α6) (Wang et al., 2007). As an electron transfer component, cyt b_5 is involved in a number of oxidative reactions, including the anabolic metabolism of fats and steroids, as well as the catabolism of xenobiotics and endogenous metabolism compounds (Schenkman and Jansson, 2003). It may serve as an obligate component or as a modifier. Cyt b_5 domains frequently fuse with other domains, such as the desaturase domain, to potentially speed up electron transfer by presenting a correctly oriented heme group, thus eliminating the need for diffusion and reorientation of the reduced cyt b_5 (Napier et al., 1997; Guillou et al., 2004). Cyt b_5 have a profound influence on cyt P450-catalyzed reactions and create a rather complex interaction with NADPH-cyt P450 reductase or NADH-cyt b_5 reductase. Cyt b_5 can receive electrons from both NADPH-cyt P450 reductase and NADH-cyt b_5 reductase, and then transfer them to different electron acceptor proteins such as cyt P450 and cyt c (Yamazaki et al., 2002; Clarke et al., 2004; Shimada et al., 2005).

Although some cyt b_5 genes have been identified and characterized in many organisms, none has been characterized in the members of the ciliate phylum, such as *Paramecium*. *Paramecium* is a well-known group of single-celled ciliate protozoa and is commonly studied as a representative of the ciliate group. *Paramecium*, which are distributed globally in a variety of habitats, play an important ecological role and often act as major trophic links in food webs (Lynn, 2008). *Paramecium* and other ciliates are the only unicellular eukaryotes that have separate germinal and somatic lineages (Jahn and Klobutcher, 2002). The germ line micronucleus is responsible for the transmission of genetic information to the next sexual generation, whereas the somatic macronucleus, containing a rearranged version of the genome streamlined for gene expression, is the seat of all transcriptional activity during vegetative growth.

The genome of *P. tetraurelia* has previously been sequenced and published (Aury et al., 2006). cDNA sequence alignments and *ab initio* predictions have estimated that the present assembly of the *P. tetraurelia* macronuclear genome sequence harbors nearly 40,000 protein-coding genes (Aury et al., 2006). This unprecedented large number of genes, especially for a unicellular organism, has turned out to be the result of at least 3 successive whole-genome duplications (WGDs). WGDs are now known to be common in the evolution of the eukaryote genome. However, these events are often resolved over evolutionary time by gene loss in the majority of the duplicates, which has previously been documented in plants, animals, and fungi (Steinke et al., 2006). In *Saccharomyces cerevisiae*, *Tetraodon nigroviridis*, or *Takifugu rubripes*,

only 12-15% of protein-coding genes evolved from duplicates formed within a WGD (Jaillon et al., 2004; Kellis et al., 2004; Brunet et al., 2006). Unlike the remnants of WGD described in other organisms, the majority of genes (68%) derived from the most recent WGD (RWGD) in *P. tetraurelia* still possess an obvious counterpart on a related chromosome, and the counterparts are always nestled between copies of the same neighboring genes. Furthermore, the RWGD coincides with the explosion of speciation events that gave rise to the *P. aurelia* complex of 15 sibling species, while the old WGD (OWGD) probably occurred before the divergence of *Paramecium* and *Tetrahymena*. In addition, due to a very low rate of large-scale chromosomal rearrangements in *P. tetraurelia*, the sequenced *Paramecium* genome provides a unique system to identify an unprecedented large number of WGD paralogs of different ages and to study the mechanisms of genome evolution consequences on gene duplication (Chen et al., 2009).

Previous studies about cyt b_5 often focused on their structures and functions (Wagner et al., 2008; Swart et al., 2011; Syed et al., 2011). However, their genome evolutionary history has not been studied so far. The availability of increasing amounts of genomic data provides us with an excellent opportunity to perform a comparative genome-wide analysis of the cyt b_5 gene family. In this study, we used comparative genomes and phylogenetic analysis to investigate the distribution of cyt b_5-like genes in the genomes of 20 species across most of the breadth of the eukaryotes. Our data showed that *P. tetraurelia* contained the largest number of cyt b_5 genes among these organisms. Taking the advantage of the unique features of the *Paramecium* genome, we performed a comprehensive set of analyses to determine the relationships among cyt b_5 genes in *Paramecium*, making significant advances in our understanding of its evolution and novel ideas about how it acquired excess genes.

MATERIAL AND METHODS

Identification of cyt b_5 genes

P. tetraurelia assembly and annotation V1.57 were downloaded from ParameciumDB (http://paramecium.cgm.cnrs-gif.fr/). A complete set of cyt b_5 sequences was identified using a reiterative process. First, the amino acid sequence of the cyt b_5 domain (Pfam: PF00173) was adopted as a query in BLASTP searches for possible homologs encoded in the *Paramecium* genome with an E-value cutoff of 1. Second, the cyt b_5 domain was confirmed by Pfam version 25.0 (http://pfam.janelia.org/) and SMART protein motif analyses (http://smart.embl-heidelberg.de/). All selected cyt b_5 genes were then used as queries to find homologs in the *Paramecium* genome by BLASTN searches. This step was crucial in allowing us to find the maximum number of candidate genes.

To exclude potentially redundant candidate cyt b_5 genes, all sequences were orientated by BLASTN, and sequences located in the same location were discarded. All non-redundant candidate cyt b_5 genes were surveyed using the Pfam and SMART databases to further verify whether they encoded cyt b_5 domains and other domains. Similar procedures were used to retrieve the cyt b_5 sequences from the other species listed in Table 1.

DNA ohnologous block alignments

In *P. tetraurelia*, the DNA-duplicated regions resulting from WGDs (called ohnol-

ogous blocks) were defined based on protein-encoding gene synteny through intragenomic comparison according to a study by Aury et al. (2006). Three successive WGDs gave rise to old, intermediate, and recent genomes, respectively. Hence, each gene from both ohnologous blocks could be assigned to their genome subsets. This yielded a complete picture of the 3 WGDs of the cyt b_5 gene in *Paramecium*.

An ohnologon was defined as a pair of ohnologous blocks that could be recognized as deriving from a common ancestral region. Ohnologous blocks were paired and ohnologons defined by comparing the identities of predicted proteins and analyzing the syntenic block. Then, the gene number in an ohnologon could be obtained and gene loss (according to a pair of ohnologous blocks, one paralog was already absent) was further analyzed (see Figure 1 for details). For example, the number of ancestral cyt b_5 genes was inferred in the ancestral genome before the 3 WGD events based on the number of genes present in the ohnologon.

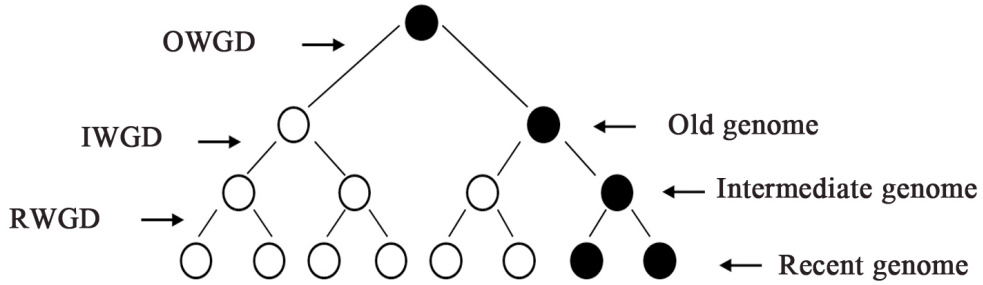

Figure 1. Successive patterns of loss in the *Paramecium* phylum. Filled circles = retained genes; open circles = lost genes. Genes from the ancestral genome gave rise to ohnolog sets of 1-2^3 members in the *P. tetraurelia* genome, depending on the timing and number of the deletion and retention events. In this example, one gene duplicate has been deleted from the old genome following the old whole-genome duplications (OWGD) and another one from the intermediate genome, following the intermediary WGD (IWGD). As a result, only 2 members descend from a unique ancestral gene after 3 successive WGDs.

Sequence alignments and phylogenetic analysis

Protein sequences were initially aligned by ClustalW. The resulting alignments were then used to guide the aligning of nucleotide coding sequences (CDSs) using MEGA v4.0 (Tamura et al., 2007). Full-length CDSs were used to generate a phylogenetic construction. Phylogenetic trees were constructed based on the bootstrap neighbor-joining method by MEGA. Support for each node was tested with 1000 bootstrap replicates.

The number of nonsynonymous substitutions per nonsynonymous site and number of synonymous substitutions per synonymous site was denoted by *K*a and *K*s, respectively. In order to detect selective constrains on duplicated cyt b_5 genes, we estimated the ratio of nonsynonymous to synonymous nucleotide substitutions (*Ka/Ks*) in the entire set of ohnologs for each WGD. *K*a and *K*s were calculated by DnaSP v4.0 (Rozas et al., 2003) based on the equations of Nei and Gojobori (1986).

Nucleotide divergence among ohnologs (the paralogs created by WGD are hereafter referred to as "ohnologs") was estimated with the Jukes and Cantor correction (Lynch and Crease, 1990) using DnaSP v4.0 (Rozas et al., 2003). Sequence exchange between duplicates was investigated by the GENECONV 1.81 program (http://www.math.wustl.edu/~sawyer/geneconv/).

A default setting with 10,000 permutations was used for the analysis, with $P < 0.05$ defined as being statistically significant.

RESULTS

Identification of cyt b_5 genes

Based on the conserved cyt b_5 domain and sequence similarity, we systematically searched cyt b_5 genes in typical species representing main lineages of the eukaryotic kingdom. In total, 310 cyt b_5 genes were identified in 20 species (Table 1). The average number of cyt b_5 genes per genome was 15.5 (ranging 6-41; Table 1). Forty-one cyt b_5 genes were found in *Paramecium*, demonstrating a significant increase when compared to other species. On the contrary, *Tetrahymena thermophila* and *Monosiga brevicollis*, the 2 protozoa most closely related to *P. tetraurelia*, contained only 13 and 11 cyt b_5 genes, respectively, and did not exhibit a similar expansion of the cyt b_5 family (Table 1).

Table 1. Cytochrome b_5 (cyt b_5) gene distribution in 20 eukaryotic genomes.

Species	Abbreviations	cyt b_5 genes	Total genes	Proportion of cyt b_5 (%)
Paramecium tetraurelia	Pte	41	39,642	0.103
Tetrahymena tetraurelia	Tt	13	24,725	0.053
Monosiga brevicollis	Mb	11	9,196	0.120
Nematostella vectensis	Nv	18	20,000	0.090
Caenorhabditis elegans	Ce	10	21,003	0.048
Drosophila melanogaster	Dm	15	14,601	0.103
Danio rerio	Dr	11	25,850	0.043
Xenopus tropicalis	Xt	9	28,000	0.032
Gallus gallus	Gg	10	21,500	0.047
Homo sapiens	Hs	14	23,224	0.060
Cyanidioschyzon merolae	Cm	6	5,331	0.113
Physcomitrella patens	Pp	20	35,938	0.056
Selaginella moellendorffii	Sm	19	22,285	0.085
Arabidopsis thaliana	At	15	27,273	0.055
Populus trichocarpa	Pt	19	30,569	0.062
Vitis vinifera	Vv	12	28,990	0.041
Oryza sativa	Os	18	37,544	0.048
Sorghum bicolor	Sb	21	35,899	0.058
Zea mays	Zm	22	53,764	0.041
Saccharomyces cerevisiae	Sc	6	6,449	0.093

In addition, the relative proportion of cyt b_5 genes to the total genes in each genome varied from 0.032 to 0.113% (0.067% on average; Table 1). Among them, 4 species had a relative proportion exceeding 0.1% (0.120% in *M. brevicollis*, 0.113% in *Cyanidioschyzon merolae*, 0.103% in *Drosophila melanogaster*, and 0.103% in *P. tetraurelia*). The high relative proportions of cyt b_5 in *M. brevicollis*, *C. merolae*, and *D. melanogaster* were possibly due to a small number of genes in their genomes (9196 in *M. brevicollis*, 5331 in *C. merolae*, and 14,601 in *D. melanogaster*). Considering the essential roles of the cyt b_5 proteins in electron transfer during oxidative reactions, the high gene proportion in small genomes might be explained by a basic requirement of the gene number. Therefore, both the high absolute number and relative proportion of cyt b_5 genes in *Paramecium* suggested that an expansion of this gene family should have occurred in the species.

Conservation and variation in domain composition and organization

Besides the cyt b_5 domain, 19 other domains were detected in the cyt b_5 family proteins, including FA desaturase domain and NAD-binding 1 domain (Figure 2). These domains formed 16 domain organizations for the cyt b_5 family in all species (Figure 2A-P). Two domain organizations existed in all species. One of them comprised the proteins that only contain the cyt b_5 domain (Figure 2A), which were predominant in all architectures (average 60.6%, ranging from 30% in *Gallus gallus* to 78% in *Oryza sativa*), suggesting that it might be a resource pool for generating novel and organism-specific domain combinations by frequent reorganization. Another domain combination (Figure 2B) consisted of the cyt b_5 and FA desaturase domains, and it was implicated in the desaturation of fatty acids. Apart from these 2 domain architectures, cyt b_5 proteins exhibited distinct sets of domain organizations among different kingdoms, suggesting that the cyt b_5 genes have undergone distinctive evolutionary paths. Plants and animals had their own specific domain organizations, namely O and C-L, respectively. By contrast, the vast majority of species in the same kingdom shared similar domain contents and organizations (Figure 3). In protozoans and animals, the domain combinations were variable and complex (Figure 3A-N), perhaps due to a complex and changing life environment. Unlike animals, plants are generally unmovable and deal with less complex surroundings, resulting in fewer and uniform domain combinations being detected (Figure 3A, B, M, and O). These results indicated that domain organizations A and B must be the most ancient of cyt b_5 protein compositions and play important and necessary roles in all species, while the other domain organizations might have been established soon after the split into different kingdoms.

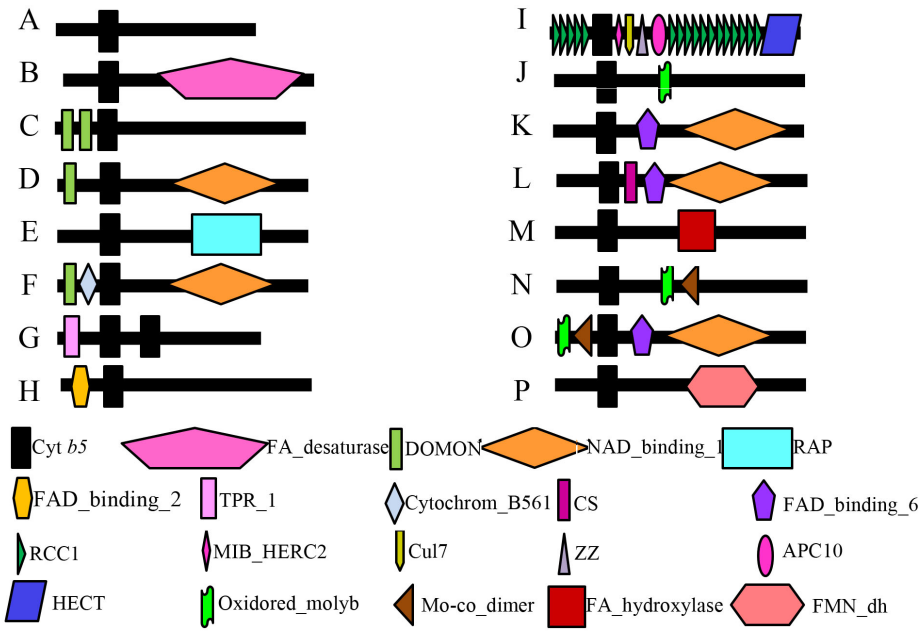

Figure 2. Domain composition and organization of cyt b_5 proteins from 20 species.

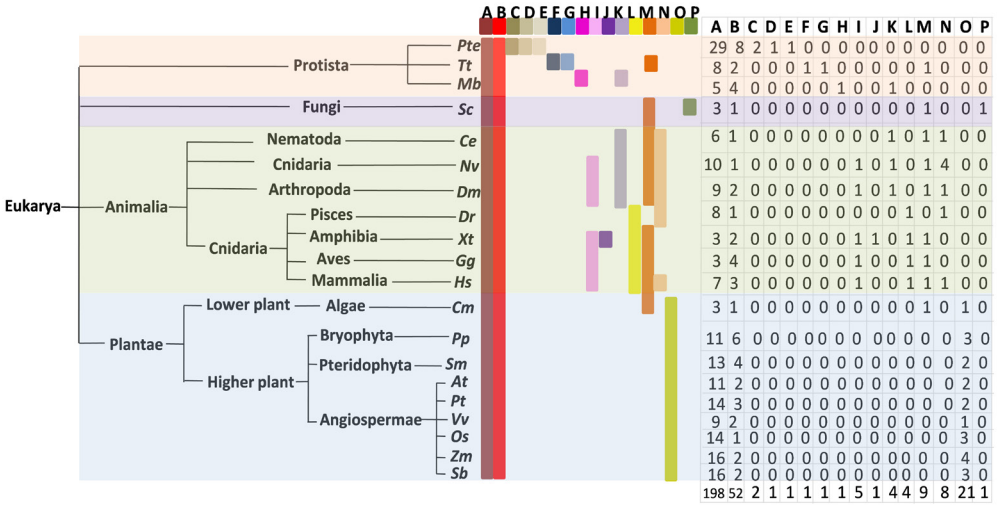

Figure 3. Distribution of domain organizations in 20 species. A-P refer to corresponding domain organizations in Figure 2. The color bars indicate that this species possesses members of this domain organization. The number of genes with different domain architectures is shown at the right.

In particular, *P. tetraurelia* and *T. thermophila* had more specific cyt b_5 protein architectures than the other species. Three of the 5 domain organizations detected in *P. tetraurelia* (Figure 3C-E) and 2 of the 5 detected in *T. thermophila* were specific to each species (Figure 3F-G), respectively. Moreover, these 2 species had several domains that were not found in other species, such as the RAP domain, which was particularly abundant in apicomplexans and might mediate a range of cellular functions through its potential interactions with RNA (Lee and Hong, 2004). This suggested that cyt b_5 proteins in protozoans might have distinct evolutionary histories.

Phylogenetic analysis of cyt b_5 genes

The above results demonstrated that the cyt b_5 gene family in *Paramecium* consisted of an extra number of genes and encoded proteins with distinct domain compositions. To analyze the evolution of this gene family in *Paramecium* further, a neighbor-joining tree was generated from the alignment of the CDSs in all investigated species (Figure S1). Most cyt b_5 genes in each species, except *P. tetraurelia*, were scattered across the branches of the phylogenetic tree. In contrast, the majority of the *Paramecium* cyt b_5 formed their own distinct branches that did not include genes from other species, suggesting that most *Paramecium* cyt b_5 had been generated by species-specific expansion. However, the bootstrap values for most branches in this tree were lower than 50%, which could not resolve a clear topology among the genes. In order to obtain a more reliable tree and to analyze the evolution of the cyt b_5 gene family in *P. tetraurelia* further, a new phylogenetic tree was reconstructed that comprised only cyt b_5 sequences from *Paramecium* and 2 functional cyt b_5 genes (Figure 4).

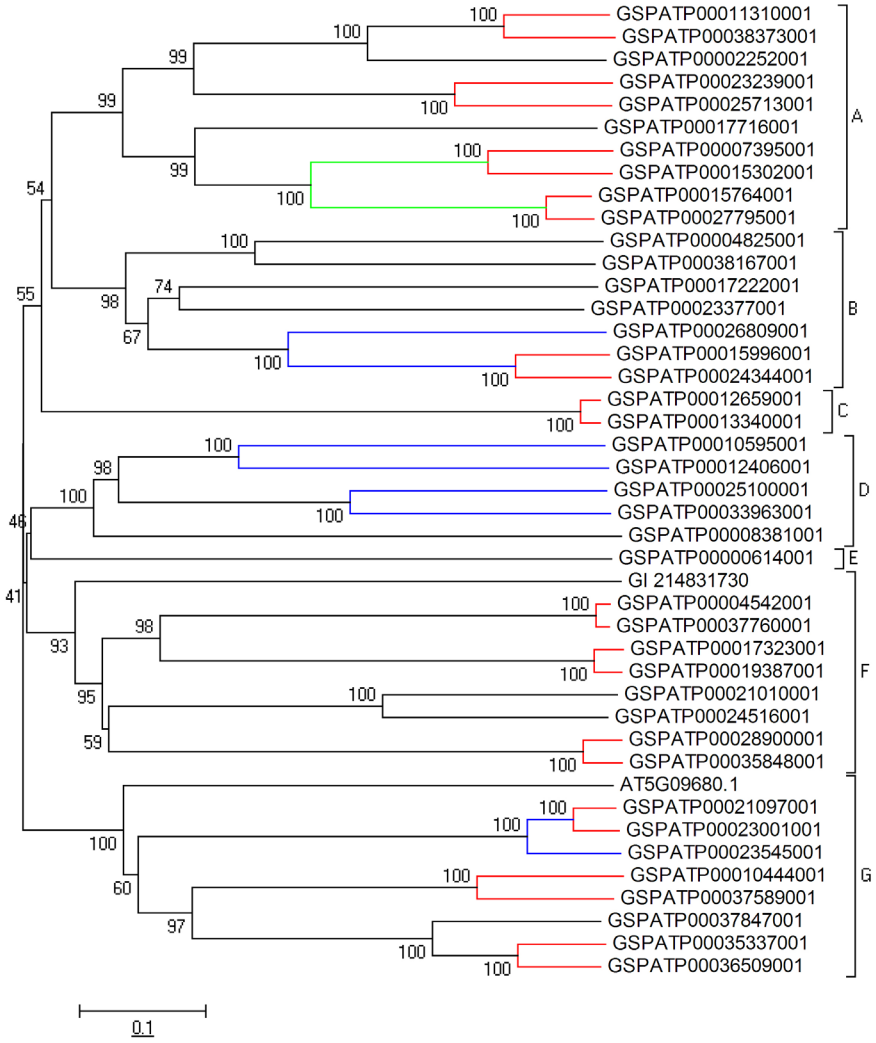

Figure 4. Phylogenetic tree for cyt b_5 in *Paramecium tetraurelia*. Colors of branches (red, blue, and green) represent cyt b_5 genes that are involved in 3 whole-genome duplications (WGDs) (recent, intermediary, and oldest WGD, respectively).

All sequences in the new tree could be divided into 6 multi-gene groups and 1 single-gene group according to the topology and bootstrap values (Figure 4A-G). Each multi-gene group contained 2 to 10 genes with similar domain compositions and organizations; and vice versa, most structurally similar genes were clustered. For instance, all 8 genes encoding proteins with FA desaturase domains were located in group F, while the proteins containing the DOMON domain were all gathered in group D. Except ohnologs, the sequences in the same group were deeply separated, indicating that either these highly divergent cyt b_5 genes were generated by a number of progenitors and/or that they had experienced a long divergence

time. The 2 functional cyt b_5 genes fell into 2 different groups. The fatty acid desaturase gene identified in humans (Njoroge et al., 2011) belonged to group F, while the other gene, which was implicated in controlling lateral root formation in *Arabidopsis thaliana* (Ikeyama et al., 2010), was a member of group G. This implied that the genes with similar structures in these 2 groups might have similar functions or might be involved in related biological pathways in *Paramecium*.

Contribution of WGD to family expansion

Phylogenetic analysis of the cyt b_5 family in *Paramecium* identified a series of paralogous gene pairs on terminal nodes of the phylogenetic tree. As has been previously shown, the *Paramecium* genome has experienced at least 3 successive WGDs and retained many duplicated sequences (Aury et al., 2006). Therefore, we examined the gene relationships inferred from a large-scale synteny analysis to determine the relative contribution of WGDs to the expansion (please refer to the Material and Methods for details). As a result, we deduced that 30 cyt b_5 genes were derived from the 3 WGDs, while the other 11 genes were orphan sequences. The RWGD explained most paralogous gene pairs, with 24 cyt b_5 copies represented in 12 two-gene pairs involved in this event, which are shown clearly by the red branches in Figure 4. The 6 genes on the left originated from the earlier WGDs. Four nodes on the tree correspond to the "intermediary" WGD (IWGD) event (Figure 4, blue nodes and branches) and 1 corresponds to the OWGD event (Figure 4, green node and branches). These findings suggested that WGDs, especially the recent one, mainly account for the species-specific expansion of the cyt b_5 gene family in the *Paramecium* genome.

To determine whether surplus cyt b_5 genes existed in *Paramecium* before the 3 WGD events, we calculated the absolute number and relative proportion of cyt b_5 genes to total protein-coding genes in its ancestral genome. By extrapolating the number of genes present in the ohnologons (pairs of ohnologous blocks that could be recognized as deriving from a common ancestral region), the absolute gene number was estimated to be 24. Compared to the numbers of cyt b_5 genes found in other species, the ancestral *Paramecium* genome still possessed the largest number of genes (Table 1). Based on the phylogenetic tree, the 24 ancestral cyt b_5 genes could be assigned to the groups A-G (group A: 5; B: 5; C: 1; D: 3; E: 1; F: 5; G: 4). According to the identification of ohnologous genes and the construction of ohnologons in the *Paramecium* genome by Aury et al. (2006), 19,552 protein-coding genes were estimated to exist in the ancestral genome before the 3 WGD events. Therefore, cyt b_5 genes accounted for approximately 0.123% of all genes in the ancestral *Paramecium* genome, which was even higher than the proportion in the current *P. tetraurelia* genome, and was the highest among all the species investigated in Table 1. Taken together, these results imply that the role of cyt b_5 genes might be so important in *Paramecium* that they were required and retained in excess both before and after the WGD events.

Nonsynonymous-to-synonymous substitution and gene conversion in cyt b_5 genes

To explore selective constrains on the duplicated cyt b_5 genes, the *Ka*/*Ks* ratio was examined for duplicates of the 3 WGDs. The *Ka*/*Ks* ratios between all duplicates were <1 and the ratio of *Ka*/*Ks* from the OWGD to the RWGD was significantly shifted toward

higher values (Table 2). The average *Ka*/*Ks* ratios between duplicates were 0.26 (ranging 0.04-0.92) in the RWGD, 0.17 (0.03-0.31) in the IWGD, and 0.16 (0.08-0.22) in the OWGD. The average *Ka*/*Ks* ratio in the recent cyt b_5 duplicates was the highest among the ohnologs and higher than that at the genome-wide level of *Paramecium* (0.08). Among the ohnologs originating from the RWGD, 2 pairs (*CytG3a*/*CytG3b* and *CytG5a*/*CytG5b*) had the highest *Ka*/*Ks* ratios (0.92 and 0.88, respectively). Thus, in general, the low *Ka*/*Ks* ratios between all duplicates indicate that purifying selection maintains the structure of the duplicate cyt b_5-like genes, whereas the increased ratio from OWGD duplicates to RWGD duplicates implies that relaxation of negative selection occurred soon after the duplication events.

Table 2. Characterization of ohnologous genes for each whole-genome duplication (WGD).

Duplication events	Number of copies	Number of genes	Average *Ka*	Average *Ks*	Average *Ka/Ks*	Average %ID (bp level)	Average %ID (aa level)
Recent WGD	2 copies	24	0.26	0.70	0.26	88.1	88.1
Intermediary WGD	2 copies	4	0.45	1.96	0.23	68.5	50.4
	3 copies	6	0.27	2.45	0.14	72.3	67.4
Old WGD	4 copies	4	0.35	2.7	0.16	67.7	53.6

The average amino acid identity of cyt b_5-like proteins in different WGDs varied from 53.6 to 88.1%, exhibiting greater diversity than that of genome-wide genes. Further evidence showed that gene conversion between duplicates was common and that it was an important evolutionary force driving gene diversification. Therefore, we utilized the GENECONV program to detect gene conversion in cyt b_5 duplicates in an attempt to identify possible evolutionary forces affecting gene diversification (Table 2). The analysis provided evidence for 4 gene conversion events among the members of this family. One was detected between a pair of ohnologs (*CytB3* exchanging with *CytB4b*) and 3 were between non-ohnologs (*CytD1b* exchanging with *CytD3*; *CytF1b* exchanging with *CytF2b*; *CytF1a* exchanging with *CytF2a*). Two exchanges (*CytF1b* exchanging with *CytF2b*; *CytF1a* exchanging with *CytF2a*) were most parsimoniously explained as an exchange between 1 common ancestor for *CytF1a*/*CytF1b* and another ancestor for *CytF2a*/*CytF2b* (Figure 4).

Exon-intron structure

Another type of variation between paralogs after duplication could be intron gain and/or loss. Therefore, we analyzed and compared the exon-intron structures of the phylogenetically closely related cyt b_5 genes in *P. tetraurelia.* Of all 41 cyt b_5-like genes detected in *P. tetraurelia*, 14 possessed no introns and the others had between 1 and 4 introns (Table 3). Here, the percentage of genes with introns (66%) was much lower than at the genome level (80%). However, the mean number of introns per gene (2.4) and the mean intron size (25.4 bp) were similar to the mean values observed in the whole *P. tetraurelia* genome (2.9 and 25 bp, respectively).

Table 3. Molecular characteristics of cyt b_5 genes in *Paramecium tetraurelia*.

Gene	Bank ID	Introns		Domain organization	%ID (bp level)	%ID (aa level)
		No.	bp			
Group A						
CytA1a	GSPATP00011310001	2	25-33	A	100	100
CytA1b	GSPATP00038373001	2	22-25	A	80.3	82.8
CytA2	GSPATP00002252001	4	23-36	E	100	100
CytA3a	GSPATP00023239001	2	22-25	A	100	100
CytA3b	GSPATP00025713001	2	22-25	A	76.5	75.1
CytA4	GSPATP00017716001	0	-	A	100	100
CytA5a	GSPATP00007395001	0	-	A	100	100
CytA5b	GSPATP00015302001	0	-	A	87.4	80.3
CytA6a	GSPATP00015764001	0	-	A	100	100
CytA6b	GSPATP00027795001	0	-	A	86.5	92.7
Group B						
CytB1a	GSPATP00004825001	0	-	A	100	100
CytB1b	GSPATP00038167001	0	-	A	41.3	45.5
CytB2a	GSPATP00017222001	2	21-24	A	100	100
CytB2b	GSPATP00023377001	0	-	A	22.7	32.9
CytB3	GSPATP00026809001	2	22-28	A	100	100
CytB4a	GSPATP00015996001	3	25-26	A	100	100
CytB4b	GSPATP00024344001	2	26	A	82.7	85
Group C						
CytC1a	GSPATP00012659001	3	24-30	A	100	100
CytC1b	GSPATP00013340001	4	19-29	A	93.9	96.9
Group D						
CytD1a	GSPATP00010595001	4	25-35	A	100	100
CytD1b	GSPATP00012406001	1	23	C	20.1	41.7
CytD2a	GSPATP00025100001	3	23-26	C	100	100
CytD2b	GSPATP00033963001	1	25	A	49.8	59
CytD3	GSPATP00008381001	2	22-25	D	100	100
Group E						
CytE1	GSPATP00000614001	2	23-25	A	100	100
Group F						
CytF1a	GSPATP00004542001	0	-	B	100	100
CytF1b	GSPATP00037760001	0	-	B	96.2	97.9
CytF2a	GSPATP00017323001	4	24-25	B	100	100
CytF2b	GSPATP00019387001	3	24-25	B	87.6	95.5
CytF3	GSPATP00021010001	1	22	B	100	100
CytF4	GSPATP00024516001	1	23	B	53.8	63.5
CytF5a	GSPATP00028900001	5	23-29	B	100	100
CytF5b	GSPATP00035848001	4	23-28	B	86.5	93.9
Group G						
CytG1a	GSPATP00021097001	0	-	A	100	100
CytG1b	GSPATP00023001001	0	-	A	86.6	93
CytG2	GSPATP00023545001	0	-	A	100	100
CytG3a	GSPATP00010444001	1	28	A	100	100
CytG3b	GSPATP00037589001	0	-	A	74.6	77.8
CytG4	GSPATP00037847001	2	26-28	A	100	100
CytG5a	GSPATP00035337001	2	22-24	A	100	100
CytG5b	GSPATP00036509001	2	22-24	A	82.8	86.5

P. tetraurelia cyt b_5 genes were designated according to their subgroup, followed by a/b standing for the ohnolog pairs inferred from whole-genome duplications (WGDs). A-E refer to corresponding domain organizations in Figure 2. %ID = identity was estimated between ohnolog pairs deriving from the recent WGD.

The 14 ohnolog pairs were divided into 2 distinct types (Types I and II), and each group comprised 7 pairs of ohnologs. There were only 2 intron numbers for Type I ohnologs (0 and 2), intron length distribution was narrow (22-25 bp), and the intron positions were identical. On the contrary, the intron number and length of ohnologs in Type II ohnologs differed from each other, but each pair shared identical positions, except *CytD1a*/*CytD1b*. For instance, the intron

positions in ohnolog pair *CytF5a/CytF5b* exhibited no change, although *CytF5b* lost an intron from its C-terminus (Figure 5). However, paralogs that were not generated from WGDs also had identical intron positions. As an example, 2 pairs of ohnologs (*CytA1a*/*CytA1b* and *CytA3a*/*CytA3b*) and a single gene, *CytA2*, which were clustered in the phylogenetic tree, shared intron positions (Figures 4 and 5). This implies that they were likely derived from the same ancestral genes and not necessarily linked to large-scale genome duplication.

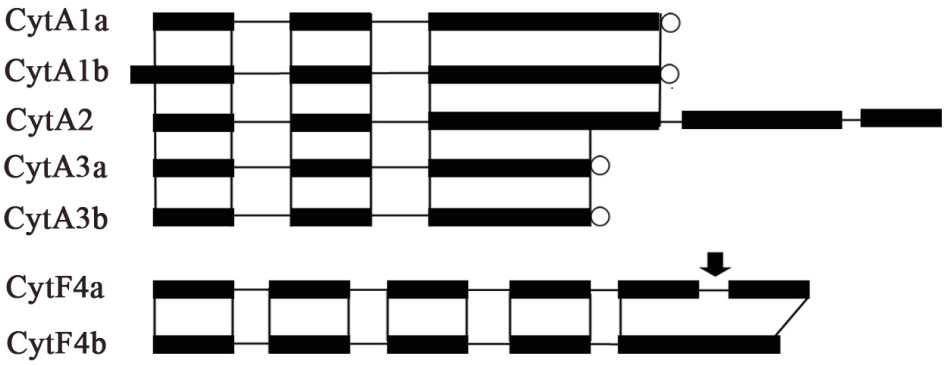

Figure 5. Exon/intron structures of duplicated cyt b_5 genes in *Paramecium tetraurelia*. Boxes and lines represent exons and introns, respectively. Arrow indicates that an intron was gained at the position, while circles indicate that introns and exons were lost.

DISCUSSION

Expansion of cyt b_5 genes in *P. tetraurelia*

As mentioned above, *P. tetraurelia* has an unusually large cyt b_5 gene family comprising 41 members. The number of cyt b_5 genes in *Paramecium* is 1.9- to 6.8-fold higher than that in other species (Table 1). The high gene number may simply be the consequence of a larger predicted gene set in *Paramecium.* However, compared with other species, a higher relative proportion of cyt b_5 genes were also presented in *Paramecium*. Furthermore, the number of cyt b_5 members in *Paramecium* was already the largest in any organism prior to the 3 WGDs. These results suggest that the cyt b_5 family in the ancestral *Paramecium* genome was already large and that further expansions occurred after the 3 WGDs, which contributed to its current size. This raises the question of why so many cyt b_5 have been maintained in *Paramecium*.

One argument is that the large number of family members exists for differential localizations and functions in the rather complex *Paramecium* cells. According to the phylogenetic tree of *Paramecium* cyt b_5, non-ohnolog sequences result in long branch lengths (Figure 4), indicating early divergence or rapid evolution. The high degree of diversity may support this hypothesis. Some studies also suggest that different members of a gene family do have specialized functions in *Paramecium*. For instance, 17 vacuolar proton-ATPases, which specifically target at least 7 different compartments, have different functions and properties in each

compartment (Wassmer et al., 2006). Twenty-six t-SNARE syntaxins that can be grouped into 15 subfamilies are assigned to different compartments and trafficking pathways (Kissmehl et al., 2007), and 16 calcineurins fulfill diverse and distinct roles in calcium-regulated processes (Fraga et al., 2010). Thus, it is tempting to speculate that the numerous cyt b_5 genes may exert site-specific functions in multiple localization sites in *Paramecium*.

Second, species-specific expansion of the cyt b_5 family allows a high degree of specialization in *Paramecium*. Unicellular organisms possess a high degree of complexity with many differentiated organelles in a single cytoplasm. In contrast, multicellular organisms have many cells and perform different functions by distributing them on different cell types with varying gene expression levels, alternative splicing, posttranscriptional modifications, and translational modifications (Kandasamy et al., 2002; Wagner et al., 2002). In *Paramecium*, these regulatory elements would have global effects that would be unable to discriminate among subcellular compartments. In addition, alternative splicing is very limited in *Paramecium*: <0.9% of the approximately 14,000 introns covered by at least 2 ESTs were found to use alternative splice sites (Jaillon et al., 2008). For example, mammalian systems increase the size of their syntaxin family by alternative splicing of members and differential expression, whereas this is not known in *Paramecium* (Kissmehl et al., 2007). Therefore, it is beneficial for *Paramecium* to retain the high number of cyt b_5 genes to increase its fitness.

Another interesting point is that the duplications of cyt b_5 genes have occurred in another 2 ancient species analyzed in this study, a moss (*Physcomitrella patens*) and a fern (*Selaginella moellendorffii*) (Figure S2). Syntenic block analysis demonstrated highly conserved genes among the flanking regions for most paralogous genes pairs at the terminal nodes of the phylogenetic trees in *S. moellendorffii* (data not shown). It illustrates that segmental duplication events are the dominant pattern for the expansion of the cyt b_5 family in *S. moellendorffii*. These findings suggest that cyt b_5 genes underwent a similar evolutionary history in ancient species, possibly as a result of adaptation to a particular environment for their survival.

Loss of cyt b_5 genes following WGDs

In general, gene loss is the main fate of duplicate genes. The *Paramecium* genes can be grouped into ohnologons of 1-2^3 members, depending on how many duplicates have been deleted/maintained following the OWGD, IWGD, and RWGD. Although the 3 WGDs may have led to ohnologons of up to 8 members, the massive loss of duplicates following the OWGD and the frequent loss of duplicates following the IWGD and RWGD resulted in reduced sets of 1 to 4 cyt b_5 ohnologs in *Paramecium*. The results of Aury et al. (2006) revealed that ~32% genes return to a single state by loss of a counterpart on a related chromosome following the RWGD (approximately 60 mya) at genome-wide level in *Paramecium*. However, as much as 54% of cyt b_5 genes were lost following the RWGD in our study, suggesting that gene loss played a major role in the current size of the cyt b_5 gene family after the duplication events. Furthermore, the current *Paramecium* genome encodes twice as many genes (~40,000 genes) than the ancestral genome (~20,000 genes) (Aury et al., 2006). In contrast, the 3 WGDs have only led to a 1.7-time increase of cyt b_5 genes, indicating that the cyt b_5 genes were present with a very limited number of duplicates in *Paramecium.* The duplication of an entire genome is considered to play a major role in evolution, and many organisms have experienced at least 1 WGD. However, a small number of genes are retained in most genomes after many rounds of

polyploidy events. Gene loss appears to have been 88% in yeast approximately 80 mya, 85% in teleost fish 300 mya, 70% in *Arabidopsis* 86 mya, and 79% in cereals 61-67 mya (Brunet et al., 2006). It suggests that increased gene copy numbers may have deleterious effects and that the dosage of these genes is under tight regulation.

Although the cyt b_5 genes were retained more frequently in *Paramecium* than in other species, they lost more duplicates than the mean level of the whole genome. It is possible that too many cyt b_5 genes would not provide extra fitness to the organism. At low cyt b_5 concentrations, cyt P450 2B4 activity is stimulated because the product is formed 10- to 100-fold faster in the presence of cyt b_5 than with cyt P450 reductase (Hlavica and Lewis, 2001). In contrast, at high concentrations, cyt b_5 inhibits the activity of cyt P450 reductase, because the binding of cyt b_5 to cyt P450 2B4 prevents the reductase from reducing the ferric cyt P450 and the catalysis process (Zhang et al., 2003). Therefore, gene loss is likely the main factor that ensures qualitative and quantitative control of cyt b_5 genes in *Paramecium.* Selective constraints on the relative dosage of cyt b_5 genes after duplication through gene retention or loss could achieve an optimum number, where the cost is balanced by the resulting benefit on fitness.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#30970198), the Key Project of the Chinese Ministry of Education (#109071), and the Qing Lan Project.

Supplementary material

REFERENCES

Aury JM, Jaillon O, Duret L, Noel B, et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444: 171-178.

Brunet FG, Roest CH, Paris M, Aury JM, et al. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* 23: 1808-1816.

Chen CL, Zhou H, Liao JY, Qu LH, et al. (2009). Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of *Paramecium tetraurelia*. *RNA* 15: 503-514.

Clarke TA, Im SC, Bidwai A and Waskell L (2004). The role of the length and sequence of the linker domain of cytochrome b_5 in stimulating cytochrome P450 2B4 catalysis. *J. Biol. Chem.* 279: 36809-36818.

Fraga D, Sehring IM, Kissmehl R, Reiss M, et al. (2010). Protein phosphatase 2B (PP2B, calcineurin) in *Paramecium*: partial characterization reveals that two members of the unusually large catalytic subunit family have distinct roles in calcium-dependent processes. *Eukaryot. Cell* 9: 1049-1063.

Guillou H, D'Andrea S, Rioux V, Barnouin R, et al. (2004). Distinct roles of endoplasmic reticulum cytochrome b_5 and fused cytochrome b_5-like domain for rat Delta6-desaturase activity. *J. Lipid. Res.* 45: 32-40.

Hlavica P and Lewis DF (2001). Allosteric phenomena in cytochrome P450-catalyzed monooxygenations. *Eur. J. Biochem.* 268: 4817-4832.

Ikeyama Y, Tasaka M and Fukaki H (2010). RLF, a cytochrome b_5-like heme/steroid binding domain protein, controls lateral root formation independently of ARF7/19-mediated auxin signaling in *Arabidopsis thaliana*. *Plant J.* 62: 865-875.

Jahn CL and Klobutcher LA (2002). Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.* 56: 489-520.

Jaillon O, Aury JM, Brunet F, Petit JL, et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.

Jaillon O, Bouhouche K, Gout JF, Aury JM, et al. (2008). Translational control of intron splicing in eukaryotes. *Nature* 451: 359-362.

Kandasamy MK, McKinney EC and Meagher RB (2002). Functional nonequivalency of actin isovariants in *Arabidopsis*. *Mol. Biol. Cell* 13: 251-261.

Kellis M, Birren BW and Lander ES (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428: 617-624.

Kissmehl R, Schilde C, Wassmer T, Danzer C, et al. (2007). Molecular identification of 26 syntaxin genes and their assignment to the different trafficking pathways in *Paramecium*. *Traffic* 8: 523-542.

Lee I and Hong W (2004). RAP - a putative RNA-binding domain. *Trends Biochem. Sci.* 29: 567-570.

Lynch M and Crease TJ (1990). The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* 7: 377-394.

Lynn DH (2008). The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature. 3rd edn. Pergamon Press, New York.

Napier JA, Sayanova O, Stobart AK and Shewry PR (1997). A new class of cytochrome b_5 fusion proteins. *Biochem. J.* 328: 717-718.

Nei M and Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-426.

Njoroge SW, Seegmiller AC, Katrangi W and Laposata M (2011). Increased Δ5- and Δ6-desaturase, cyclooxygenase-2, and lipoxygenase-5 expression and activity are associated with fatty acid and eicosanoid changes in cystic fibrosis. *Biochim. Biophys. Acta* 1811: 431-440.

Ogishima T, Kinoshita JY, Mitani F, Suematsu M, et al. (2003). Identification of outer mitochondrial membrane cytochrome b_5 as a modulator for androgen synthesis in Leydig cells. *J. Biol. Chem.* 278: 21204-21211.

Rozas J, Sanchez-DelBarrio JC, Messeguer X and Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.

Schenkman JB and Jansson I (2003). The many roles of cytochrome b_5. *Pharmacol. Ther.* 97: 139-152.

Shimada T, Mernaugh RL and Guengerich FP (2005). Interactions of mammalian cytochrome P450, NADPH-cytochrome P450 reductase, and cytochrome b_5 enzymes. *Arch. Biochem. Biophys.* 435: 207-216.

Soucy P and Luu-The V (2002). Assessment of the ability of type 2 cytochrome b_5 to modulate 17,20-lyase activity of human P450c17. *J. Steroid Biochem. Mol. Biol.* 80: 71-75.

Steinke D, Hoegg S, Brinkmann H and Meyer A (2006). Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol.* 4: 16.

Swart P, Goosen P, Storbeck KH and Swart AC (2011). A regulatory role for cytochrome b_5 (Cyt b_5) in adrenal steroidogenesis: influence on 3 beta-hydroxysteroid dehydrogenase-Delta 5-isomerase (3 beta HSD) activity. *FEBS J.* 278: 188-189.

Syed K, Kattamuri C, Thompson TB and Yadav JS (2011). Cytochrome b_5 reductase-cytochrome b_5 as an active P450 redox enzyme system in *Phanerochaete chrysosporium*: atypical properties and *in vivo* evidence of electron transfer capability to CYP63A2. *Arch. Biochem. Biophys.* 509: 26-32.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.

Vergeres G and Waskell L (1995). Cytochrome b_5, its functions, structure and membrane topology. *Biochimie* 77: 604-620.

Wagner CR, Mahowald AP and Miller KG (2002). One of the two cytoplasmic actin isoforms in *Drosophila* is essential. *Proc. Natl. Acad. Sci. U. S. A.* 99: 8037-8042.

Wagner S, Kiupel M, Peterson RA, Heikinheimo M, et al. (2008). Cytochrome b_5 expression in gonadectomy-induced adrenocortical neoplasms of the domestic ferret (*Mustela putorius furo*). *Vet. Pathol.* 45: 439-442.

Wang L, Cowley AB, Terzyan S, Zhang X, et al. (2007). Comparison of cytochromes b_5 from insects and vertebrates. *Proteins* 67: 293-304.

Wassmer T, Kissmehl R, Cohen J and Plattner H (2006). Seventeen a-subunit isoforms of paramecium V-ATPase provide high specialization in localization and function. *Mol. Biol. Cell* 17: 917-930.

Yamazaki H, Nakamura M, Komatsu T, Ohyama K, et al. (2002). Roles of NADPH-P450 reductase and apo- and holo-cytochrome b_5 on xenobiotic oxidations catalyzed by 12 recombinant human cytochrome P450s expressed in membranes of *Escherichia coli*. *Protein Expr. Purif.* 24: 329-337.

Zhang H, Gruenke L, Arscott D, Shen A, et al. (2003). Determination of the rate of reduction of oxyferrous cytochrome P450 2B4 by 5-deazariboflavin adenine dinucleotide T491V cytochrome P450 reductase. *Biochemistry* 42: 11594-11603.